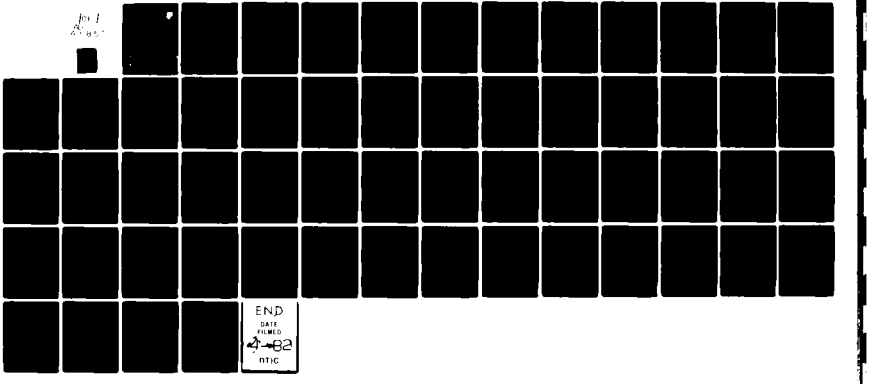
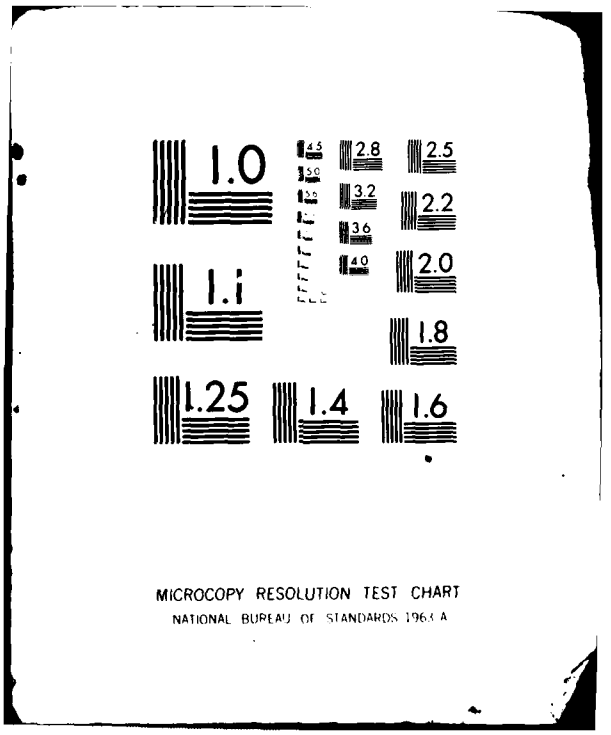
AD-A111 857 VERBEX CORP BEDFORD MA F/G 5/7
EXISTING TECHNIQUE DEVELOPMENT.(U)
DEC 81 P G BAMBERG, L G BAHLER, J M BAKER F30602-80-C-0203
UNCLASSIFIED RADC-TR-81-355 NL



END
DATE FILMED
4-82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

(12)

RADC-TR-81-355
Final Technical Report
December 1981

GISTING TECHNIQUE DEVELOPMENT

Verbex

Paul G. Bamberg
Larry G. Bahler
Janet M. Baker
Henry G. Kellett

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-81-355 has been reviewed and is approved for publication.

APPROVED: Joseph T. Nelson

JOSEPH T. NELSON, 1LT, USAF
Project Engineer

APPROVED: John N. Entzminger Jr.

JOHN N. ENTZMINGER, JR.
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-81-355	2. GOVT ACCESSION NO. AD-A111 857	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) GISTING TECHNIQUE DEVELOPMENT		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report 10 Jun 80 - 9 Jun 81
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Paul G. Bamberg Henry G. Kellett Larry G. Bahler Janet M. Baker		8. CONTRACT OR GRANT NUMBER(s) F30602-80-C-0203
9. PERFORMING ORGANIZATION NAME AND ADDRESS Verbex 2 Oak Park Bedford MA 01730		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 31011G 70550735
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAA) Griffiss AFB NY 13441		12. REPORT DATE December 1981 13. NUMBER OF PAGES 60
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Joseph T. Nelson, 1Lt, USAF (IRAA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Wordspotting
Keyword Recognition
Connected Speech Recognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report documents the methods utilized to improve and simplify the procedure for operating reference templates and word models used in the keyword recognition process. Commands necessary for the automatic generation of reference templates have been added and the procedure for word model generation has been automated. Test results show a modest performance improvement over previous methods. Recognition was improved with a 20-word English set from 33.5% to 41% operating at a threshold of 2.52 false

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

alarms/hr/word. Techniques have also been developed for on-line reference generation that requires no auxiliary mass storage devices. These techniques are also described.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

Introduction and Summary

Verbex has worked on four previous keyword spotting contracts from RADC, the most recent being Contract F30602-78-C-0036. (1). These contracts had as test material high fidelity speech input, small vocabularies, and scripts that were read. The results of these efforts produced a recognition rate of 85% detection probability at 5 false alarms per hour for speaker independent continuous speech on the word "Kissinger". In the most recent contract work (1), we began efforts at automatic template generation and a 20 word vocabulary of words active simultaneously for either Russian or English speech. Typical results for males were: a recognition rate of 33.5% for 20 English words at 2.52 false alarms per hour; for females a 19% recognition rate for 20 Russian words at 5.7 false alarms per hour.

The work performed in fulfillment of the present contract, Gisting Technique Development (NO. F30602-80-C-0203), is a direct consequence of the work described in the previous report.(1) In the previous contract we had made an initial effort at increasing capability of the system by going from high fidelity to telephone speech input, from a one or two word vocabulary to a twenty word vocabulary, and testing on conversational speech. In this contract we continued this effort. In it work has focused on improving and simplifying the procedure for generating the reference templates and word models used in the keyword recognition process. The signal processing and dynamic programming techniques described in the previous

report have been left essentially unchanged, and the same database has been used for template generation and for performance testing. We have added to the keyword system all the commands necessary for automatic generation of reference templates on an existing database of labeled keywords, and we have automated the procedure for generating word models. Test results indicate a modest improvement in recognition performance over what was obtained with previous reference templates and word models. For example, the recognition rate for males on 20 English words went from 33.5% to 41% with 2.52 false alarms per hour. This work is described in Chapter II of the report.

As a tentative further step in automating the generation of reference templates and word models, we have experimented with on-line generation of templates and word models directly from microphone input to the system. The techniques we have developed require no auxiliary mass storage devices and call for only minimal skill on the part of the operator in interpreting acoustic displays. These on-line techniques have been tested only informally, in a speaker-dependent mode, but the preliminary results are encouraging: an inexperienced operator was able, in a few hours' work, to train a complete new vocabulary of twelve words which appears to recognize at least as well as existing speaker-independent templates recognize the standard twenty-word vocabulary. This work is described in Chapter III of the report.

(1) S.L. Moshier and L.G. Bahler, J.M. Baker, "Keyword Operational Analysis", 26 May 1981.

The Statement of Work calls also for recognition of both male and female speakers. We have added this capability to the system in two different ways. First, we have permitted the storage of two sets of reference patterns, one for males, one for females, between which the operator can transfer freely and rapidly. Second, we have conducted extensive research into frequency-warping techniques which permit patterns generated from speakers of one sex to be used in recognizing speakers of the other sex. These techniques were developed and thoroughly tested in the context of isolated-word recognition, and many of them now have been transferred to the key-word recognition system. Our tests indicate that frequency warping can convert male reference patterns which are quite inadequate for recognizing female speakers into patterns whose performance is not too far inferior to the performance of patterns generated from female speakers. This work is described in Chapter IV of the report.

We have tested the reference patterns and word models generated by the newly developed techniques on the same government-furnished tapes ("Stonehenge" database) which were used for previous contracts. Recognition results for English male and female speakers are presented in Chapter V of the report. For quality 2 noise conditions with 20 words, recognition for males was 41% at 2.5 false alarms per hour; for quality 3 noise recognition was 32% at 2.9 false alarms per hou. For females the results were about 10% less for about twice the false alarm rate. Tables are not presented for Russian speakers, as there was no improvement over the results of the previous contract (1).

CHAPTER II

Automatic Generation of Reference Patterns and Word Arrays

A. Strategy for Reference Pattern and Word Array Generation

Associated with each vocabulary word in the keyword system are two data structures, a word array and a set of reference patterns. The word array specifies the number of segments into which the word is divided, the name of the reference pattern associated with each segment, and the minimum and maximum allowed duration for each segment. Each reference pattern incorporates the means and standard deviations of the acoustic parameters associated with one segment of a word. During recognition, acoustic parameters from each 10-millisecond frame of speech are compared with all reference patterns to yield a set of pattern scores. The word array then controls the conversion of these pattern scores into word scores by the dynamic programming algorithm described in the previous report (1). The effect of the dynamic programming is to find the best match between the recent speech input and the sequence of patterns for each word. If this is good enough for a particular word, detection of that word is announced. At this point the algorithm has associated each input frame with a segment of the word, in compliance with the constraints imposed by the word array.

Speaker-independent reference patterns and word arrays are generated from the design database of labeled, digitized speech waveform data described in the previous report (1). The English male database, for example, contains approximately a hundred instances of each vocabulary word, spoken by 28 different speakers. The identity, location, and duration of

each occurrence of a vocabulary word in the database are recorded in a separate label file. Reference patterns and word arrays are produced by making one pass through the label file alone, followed by two passes through the entire database as follows:

Pass 1: From the labels alone, the mean duration of each vocabulary word is calculated. From these mean durations, the operator generates an appropriate "open word array" which specifies the number of patterns for each word and sets a resonable minimum and maximum duration for each. To permit direct comparison with previous results, our word arrays have used the traditional eight patterns per word. For each segment, we allow a minimum duration of one frame and a maximum duration of one-fifth of the mean total duration of the word.

Pass 2. Using the open word arrays, the system generates a list of the reference patterns which must be constructed and assigns an identifying number to each. It then proceeds through all the design data, carrying out the signal processing to generate acoustic parameters. On reaching the end of each labeled word, it selects as many sample frames, uniformly spaced between the beginning and end of the word, as there are patterns for that word. For example, if a word has eight patterns and a labeled instance of it extends over 42 frames, frame 7 (the first frame whose parameters are all calculated from data within the labeled utterance) will be assigned to pattern 1 for the word, and frames 12, 17, 22, 27, 32, 37, and 42 will be assigned to patterns 2 through 8 respectively. The sums and sums of

squares of parameters for the frames thus assigned to each pattern of each word are accumulated. At the end of the pass, means and standard deviations are computed from these accumulated data to create initial reference patterns for each word. The effect of this procedure is that, for example, the last pattern for a word incorporates the means and standard deviations of all parameters for the last frame in each labeled instance of the word.

Pass 3: The recognition algorithm is run for all the design data, using the open word arrays generated in pass 1 and the initial reference patterns generated in pass 2. Whenever the end of a labeled word is reached, a traceback is performed to dtermine whether the word was recognized with a hypothesized starting frame within 10 frames of the labeled start. If so, the number of frames assigned by the dynamic programming to each pattern in the word is recorded in a histogram which will be used in constructing an improved word array, while a frame chosen from near the center of the sequence of frames assigned to each pattern is accumulated for use in making new reference patterns.

Occasionally the hypothesized and labeled start of a word differs by more than ten frames. This indicates that the open array and initial reference patterns were not good enough to permit dynamic programming to match the labeled instance of the word to the word model. In such a case, no data is accumulated. As a result, not all the design data are used in constructing the final reference patterns. For most keywords, though, the fraction of unused instances is small.

At the end of this pass, final refernce patterns are made from the accumulated sample frames, while a final word array is made from the histogram of segment duration data. The reference patterns and word array incorporate data from all the labeled utterances which were satisfactorily recognized.

This batch-processing strategy is fundamentally the same as what has been done in the past, but it is faster, more automatic, and has led to better recognition. Here is a list of some of the major improvements.

1. All data storage is now in the vector processor memory, and all code is now part of the keyword software. Previously, sample frames were written to external files and processed by other programs.
2. The open word arrays, which take advantage of knowledge of the mean duration of each word, lead to a higher percentage of correct recognitions during the final pass, so that a higher percentage of the design data is used in constructing the final patterns and word arrays.
3. Construction of the open word arrays requires the operator to insert only one parameter per word into a file; construction of the final word arrays is completely automatic. Previously, making word arrays required inserting duration data for each pattern into an assembly-language program, then assembling and linking this program.

4. The computation for the reference patterns is now performed entirely in the vector processor and requires only a second or two once the sample frames have been accumulated. The previous algorithm took several minutes to achieve the same result.

B. New Data Structures and Procedures

To implement the new automatic procedures, several new data structures and many new commands have been added to the system. The following is a brief summary of these.

1. Word arrays

There are two word array files, a primary file and a secondary file, in vector processor memory. The secondary file is built up by the operator; it may include copies of word arrays for exisiting words taken from the primary file, and open word arrays for new words. Either file may be used to define the current vocabulary.

2. Word Durations

Data on the durations of all vocabulary words may be accumulated in the host computer memory. The operator can request a printout of the mean and standard deviation of the durations for each vocabulary word for guidance in constructing open word arrays.

3. Segment Durations

As words are recognized, a histogram of the measured duration of each segment is accumulated in the vector processor memory. The operator can

establish the minimum and maximum allowable duration for each segment in the final word array either as the 25th and 75th percentiles from this histogram or in terms of the mean and standard deviation of the measured durations for the segment.

4. Accumulated Sample Frames

Accumulated sample frames are stored in a file in the vector processor memory. At the operator's command, new reference patterns are created for all words for which there are data in this file.

5. Reference Patterns

There are two reference pattern files in vector processor memory. When changes are being made to the vocabulary, patterns for existing words may be copied from one file to the other while patterns for new words are being computed from accumulated sample frames.

C. Summary of Test Results

The reference patterns and word arrays generated by the automatic procedures described above were tested by using them in recognizing both the design data from which they were produced and the independent "Stonehenge" test data. A variety of options in generating word arrays were tried; the results below describe the most successful of these.

The test results report the percentage of detections for the entire vocabulary of 20 words when the detection thresholds for individual words were set to restrict total false alarms per word to 5, 10, 15 or 20 respectively over

the entire test set. The duration of the design database is 0.92 hours, while that of the test database is 1.19 hours. The four tests for English males used the following reference patterns and word arrays.

Test 1: Best previous results, using reference patterns generated by the standard Verbex statistics program and a hand-made word array based on selection of the 25th percentile as minimum duration, the 75th percentile as maximum duration.

Test 2: Reference patterns were the initial reference patterns generated by the keyword system, and word arrays were generated automatically by the keyword system using 25th and 75th percentiles.

Test 3: Reference patterns were again the initial reference patterns generated by the keyword system; word arrays were generated automatically, using mean $\pm$ 0.75 standard deviations as maximum and minimum durations.

Test 4: Word arrays were same as in test 3, but with final reference patterns generated during the pass that produced the word array, using a sample frame from the center of each identified segment of the word.

		False Alarms Per Word			
		5	10	15	20
Test 1	Design Data	60%	68%	72%	74%
	Test Data	39%	48%	51%	55%
Test 2	Design Data	61%	69%	72%	75%
	Test Data	43%	49%	54%	57%
Test 3	Design Data	61%	68%	72%	74%
	Test Data	42%	47%	54%	57%
Test 4	Design Data	63%	70%	74%	77%
	Test Data	47%	53%	57%	59%

Table I: Test Results Using Different Methods of Reference Pattern Generation

Complete results for Test 4 on a word-by-word basis are presented in Chapter V, along with comparable data for English female speakers.

A similar test was performed also for Russian female speakers. Since the duration of the database is different, the results cannot be compared directly with those for English, but they do permit an assessment of the automatic word-array generation procedure.

		False Alarms Per Word			
		5	10	15	20
Test 1 (best previous)	Design Data	38%	45%	50%	53%
	Test Data	26%	30%	33%	39%
Test 2 (automatic word array-25th-75th percentiles)	Design Data	42%	50%	55%	59%
	Test Data	27%	33%	38%	41%

Table 2: Test Results for Russian Keywords Using Different Reference Generation Methods.

The improvement in recognition reflected in these test results is to be credited not to the initial reference patterns, which were essentially identical to previous ones, but to the new word arrays and to the final-pass reference patterns used in Test 4 of the English data. Since our initial aim was simply to duplicate the existing word arrays without the need for manual intervention, the improvement comes as a pleasant surprise. It probably results not from the automatic procedures for extracting percentiles from histograms, but from the improved initial word arrays which were used during the final pass through the data, in which the segment durations were measured. By constraining the durations more tightly, and in a manner appropriate to the length of each word, these arrays permitted a high percentage of the

labeled keywords to be recognized satisfactorily, thereby increasing the quantity of the data that contributed to the final word arrays.

D. Suggestions for Further Research

The new commands added to the system in order to generate word arrays automatically have opened up several lines of research for which the software development is essentially complete, and the remaining task consists primarily of testing on the full database. These include the following:

1. Iteration of the final pass to generate still better word arrays and reference patterns. Preliminary efforts in this direction have produced no further improvement, but there are many parameters to be varied, and we anticipate that it is possible to design a multi-pass procedure which ultimately converges to a final "best" word array and "best" set of reference patterns for a given set of design data.

2. Generation of final reference patterns and word arrays in only a single complete pass through the data. The idea is to generate initial reference patterns from a subset of the design data, then to generate word arrays and final reference patterns iteratively as the data are accumulated, using the sort of techniques that will be described in the next section.

3. Use of word-duration data to constrain maximum and minimum total word duration. This capability has been present in the system for a long time, but only recently have the statistics on word duration become available. It is possible that some

of the present false alarms are the result of "recognitions" in which all the individual segment durations are reasonable, but in which the sum of all segment durations exceeds the mean measured duration of the word by several standard deviations. Such false alarms could be eliminated by a well-designed constraint on maximum word duration.

4. Using a variable number of segments per word. At present, all vocabulary words, from "look" to "Westchester", have eight patterns associated with them. In view of our notable lack of success to date in detecting one-syllable keywords, there is nothing to lose by trying word arrays in which shorter words have fewer patterns. Since memory capacity and computation speed restrict the number of patterns, not the number of vocabulary words, it might prove possible to enlarge the vocabulary without degrading recognition. One ultimate aim of this research would be automatic generation of word arrays which have the optimal number of segments for each word.

5. Multiple attempts to recognize difficult words. Our experiments with isolated word recognition suggest that design data which is difficult to recognize contributes more to speaker-independent reference patterns than data which is easy to recognize. Such data cannot always be used in the present keyword system, because the open word arrays used during the final pass through the design database do not always generate an acceptable match of segments to each labeled word. Given the present capability of having two alternate word arrays, it would not be difficult to make two attempts to recognize such words, using two different open word arrays.

CHAPTER III

III. On-line Training for Speaker-Dependent Keyword Recognition

A. Description of Training Technique

Using the system commands developed for automatic generation of reference patterns and word arrays in the traditional batch-processing procedure, we have developed a preliminary version of a technique for generating patterns and word arrays from direct voice or tape input. This procedure needs refinement and testing, but already it seems to work fairly well for speaker-dependent recognition. It requires no mass storage devices ; everything is stored in the vector processor memory. All the operator must do is, in the initial phase, label the start and end of isolated words on a display of amplitude as a function of time, then, in the later phases, indicate whether the recognized word was what he spoke.

Here are the instructions for training a single new word. The command codes are included only to indicate how much typing is involved; it is not necessary to be familiar with all the system commands in order to carry out the procedure. For the sake of concreteness, we assume that the operator is training a new word 3 in the vocabulary, and that words 1 and 2 have already been trained by this procedure.

Step 1: Create an open word array for the new word.

Commands: SA - switch to secondary word array file

908,3NW* make a new open word array for word 3, with 8 patterns, each having a required duration of 1 frame, an optional duration of 9 frames, for a total allowed maximum duration of 10 frames

PA - switch to primary word array file

3UW - copy the new array for word 3 to the file

3ZA - clear out accumulators for word 3

IH - initialize duration histograms to zero

ZL - clear word duration accumulators

Step 2: Generate initial patterns from isolated utterances

Commands: DSKN - enable XY-display and cursor knob

0TH - set thresholds to zero to disable recognition

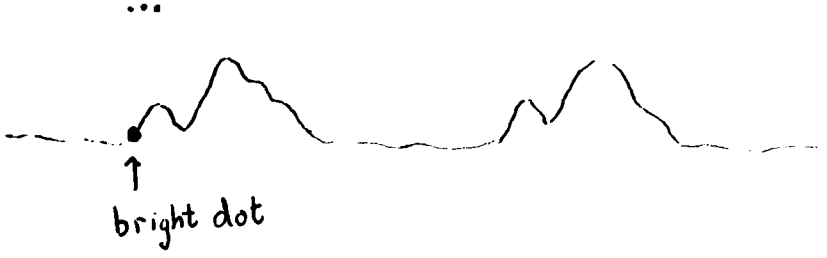
GO - begin processing live input

Now, after a second or two of silence, speak the new word twice, with a pause of about ½ second between utterances. Allow about a half-second of silence, then strike any key on the terminal. The lowest display will show amplitude as a function of time for the last 2½ seconds, for example:

*. (100I + J), KNW means "create a model for word K with J patterns, each having a required duration of 1 frame, and an optional duration of I frames."

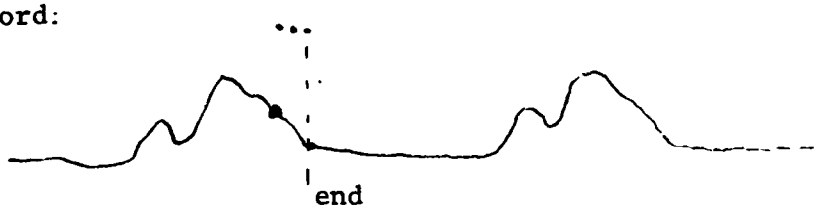
If two nearly identical patterns stand out from the silence, each must be the desired word. Otherwise, type "GO" and try again.

Once the patterns are found, turn the cursor knob until the bright dot in the amplitude display is at the left edge of the left utterance:



Command: S - mark start of word

Now move the bright dot to seven frames from the end of the word, so that the right-hand one of the three dots above the display matches the end of the word:



Command:	3EA	-executes the following sequence:
	E	-mark end of word
	3WD	-label word as number 3
	AC	-accumulate equally spaced frames from word
	XL	-accumulate labeled word duration

Repeat this labeling procedure for the second

utterance, then repeat the entire procedure of speaking and labeling two words four or five more times. At this point it is possible to generate some initial reference patterns.

Command: MR -generate reference patterns from accumulated sample frames.

VL -calculate mean and standard deviation for labeled word durations and display them at terminal.

Step 3: Recognize isolated words

Commands: 50UA - set to take sample frame from middle of each segment

20TH - set recognition threshold to 20. This number should be lowered if there are too many false alarms,raised if there are too many missed detections.

4GO - recognize live input, stop on detection and type length of word.

Now speak the vocabulary word. Repeat it, if necessary, until the detection symbol "C" (third word) appears on the screen. The number following the symbol is the length of the word as recognized. If this is within two standard deviations of the mean duration, and if the detection symbol was printed shortly after the word was spoken, accept the word as recognized.

Command: OK -executes the following sequence:

-3CE - move end of label back 3 frames to correct for tendency of automatic labeling to go beyond end of word.

-UH - record segment durations and accumulate sample frame

XL - accumulate labeled word duration

MR - remake reference patterns

4GO - continue recognizing, stop on next detection

If an incorrect detection symbol or unreasonable word length appears, type "4GO" instead of "OK," and the detection will be ignored.

Repeat this procedure until five or six "OK" words have been detected.

Step 4: Generate a word array:

Commands: 150,3SH - select maximum and minimum durations as mean $\pm$ 1.5 standard deviations

3RH - make array for word 3 from histogram of durations

3UW - write this new word array to file

Step 5: Recognize words in context

The procedure is the same as in step 3, except that now, after typing "4GO" or "OK", the new word should be spoken in connected speech, in a variety of contexts. After a few successful recognitions, step 4 should be carried out again, but with "75, 3SH" as the first command. This will create a better-constrained word array.

Step 6: Test the new patterns, set the threshold

Commands: 3,10ST -set threshold for word 3 to 10

GO -start recognition

Now read a text containing several instances of the new word. Note whether missed detections or false alarms are a more serious problem. Hit any key on the terminal to stop recognition. If false alarms are a problem, type

3,2DT - decrease threshold for word 3 by 2

GO - resume testing

If missed detections are a problem, type

3,2IT - increase threshold for word 3 by 2

GO - resume testing

Once a reasonable threshold has been found, the patterns should be fairly satisfactory. They can be improved further, if desired, by repeating steps 5 and 6.

B. Preliminary Results

One of the authors (PGB), who has not learned to interpret any of the displayed data except the amplitude data mentioned in Step 2, has used the above procedure to train the following 12-word vocabulary, chosen from a computer language reference manual.

1. identifier
2. expression
3. operator
4. function
5. arithmetic

6. pointer
7. assignment
8. constant
9. value
10. integer
11. character
12. type

The training procedure required about four hours, with the last few words requiring ten or fifteen minutes each to train. The resulting speaker-dependent patterns and word arrays performed quite satisfactorily in recognizing, for example, most of the keywords in sentences like "The value of an arithmetic expression may be an integer but not a character." Since speaker-dependent recognition is generally a much simpler task than speaker-independent recognition, this early success is not surprising, nor is it surprising that the author's patterns performed much less well in recognizing a different male speaker.

Our tentative conclusions are that on-line training of speaker-dependent patterns from a cooperative speaker is feasible on a system with no mass storage devices attached, that an experienced operator can quickly learn to train new words at the rate of four or five per hour, and that the resulting patterns frequently perform as well for the speaker who made them as do speaker-independent patterns made from a large, labeled database. While we have not yet had time to carry out any quantitative tests,

even the obvious one of retraining the "Stonehenge" English vocabulary on-line, we feel optimistic about prospects for further development of on-line training.

CHAPTER IV

IV Speaker Normalization

a. General Strategy

Characteristically, a speaker-independent recognition system attempts to average together acoustic parameters for a wide variety of speakers in making reference patterns, with the hope that the resulting patterns will perform satisfactorily in recognizing any speakers similar to those in the design data. In practice, this approach leaves something to be desired; for example, reference patterns made from 90% male speakers,10% female speakers do not perform nearly as well in recognizing females as in recognizing males.

One approach to improving speaker-independent recognition is to have alternate sets of reference patterns, one for males, one for females, for example, or one for each of several dialects. This approach has been implemented successfully in the Verbex speaker-independent isolated-word recognition system by means of clustering algorithms which automatically select an optimal set of alternate patterns for each word from a given database. The price one pays is a significant increase in computation, since likelihood scores must be computed for all alternate patterns, and in memory requirements, since alternate patterns for each word must be stored.

For the keyword system, we have pursued a different approach: speaker normalization. Here the idea is to develop a transformation for each speaker which converts the acoustic parameters for that speaker's utterances into the corresponding parameters for the same word as spoken by a "standard" speaker or, equivalently, which converts reference patterns appropriate to a "standard" speaker into patterns appropriate to one particular speaker. There are two aspects to implementing this approach: first, computing the transformed acoustic parameters, second, devising an appropriate transformation for each speaker on the basis of a limited sample of speech. Both aspects have been thoroughly tested in the context of isolated words, but only the first has yet been transferred to the keyword system.

B. The Cubic Spline Algorithm

For each 10-millisecond frame of speech, the signal processing algorithm generates 31 acoustic parameters which are determined by the power in the input signal at 31 different frequencies, which will be denoted $f_1, f_2 \ldots f_{31}$. We may think of these parameters as sampling a function P(f) at 31 different frequencies.

To implement speaker normalization, one must sample the function P at a different set of frequencies, $f^*_1, f^*_2, \ldots f^*_{31}$, chosen so that the parameter $P(f^*_k)$ for the actual speaker is optimally matched to the reference-pattern parameter for the "standard" speaker at frequency f_k. For example,

if a given female speaker has a vocal tract 15% shorter than a "standard" male speaker, one might expect the acoustic parameter for the female speaker at a frequency of 1150 Hz to correspond to the 1000 Hz acoustic parameter for the "standard" speaker.

The mathematical problem is simply one of interpolation, as illustrated in Figure 2 below. We know the values of the function P (f) at 31 specific frequencies and wish to estimate the value of this function at other frequencies. The solution to the problem is suggested by the graph: draw a"smooth curve" through the given function values (solid dots), then use this curve to estimate the value of the function at other points. (crosses).

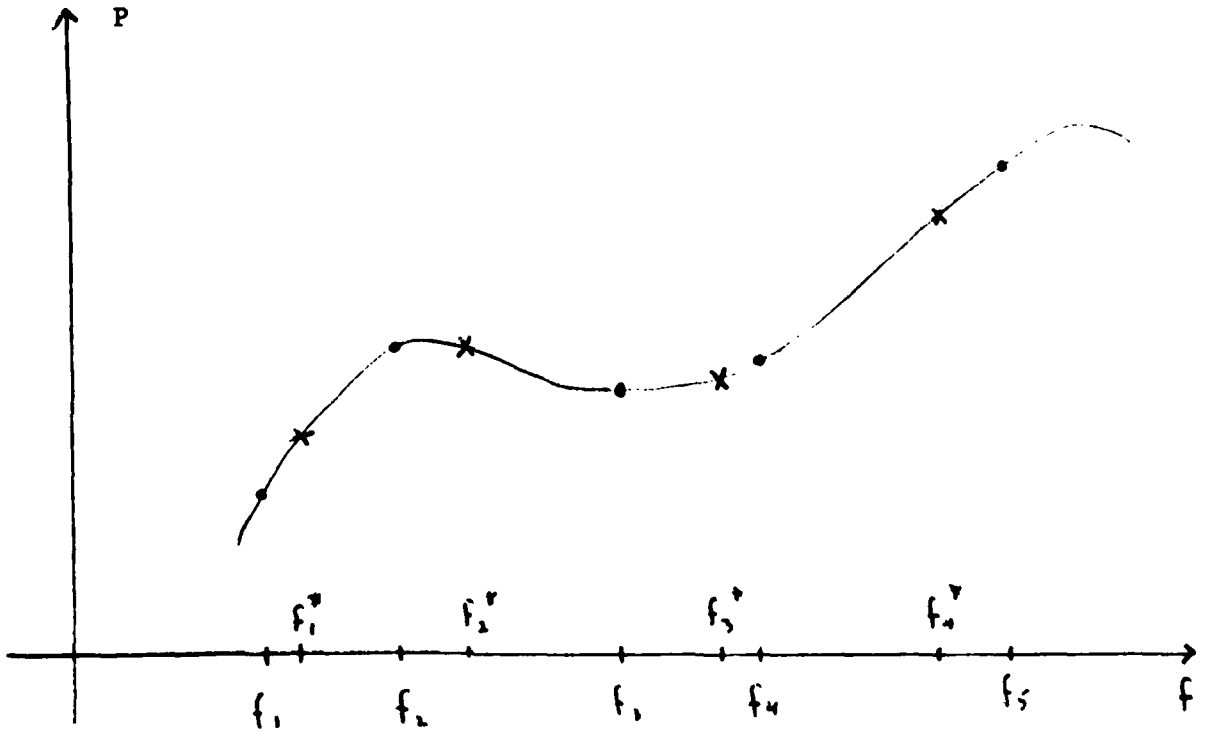


Figure 2 Cubic Spline Interpolation

The "smooth curve" is constructed by an algorithm which mimics the behavior of a draftsman's spline, a flexible bar with weights that are placed atop all data points. Such a spline assumes a shape that has the minimum possible total curvature of any curve that passes through the given data points. The curve constructed by our cubic spline algorithm has the following equivalent set of properties.

1. It passes through all data points.
2. Between any two data points, it is a cubic polynomial.
3. At the data points, these cubic segments join as smoothly as possible: the curve is continuous, and so are its first and second derivatives. The resulting curve has continuous slope and curvature and appears to the eye to be perfectly smooth.

These properties, supplemented by a condition on the third derivative of the function at the endpoints, determine the interpolating curve completely. To calculate the parameters for each cubic segment, it is necessary to solve a system of linear equations, one per data point. To evaluate the interpolating function, all that is required is evaluation of a cubic polynomial. What makes the spline algorithm attractive for the application at hand is that, unlike many other interpolation algorithms, it requires neither equally spaced function values nor assumptions about the overall form of the interpolating function.

Initially the spline algorithm was written for the VAX in a higher-level language using floating point arithmetic.

It proved effective but slow. Once thoroughly tested, it was recoded for the vector processor, using integer arithmetic. The present version of the algorithm is almost free of round-off error for the 8-bit acoustic parameters we use, yet it is so fast that to transform a complete set of 162 reference patterns, for each of which three interpolating curves must be calculated, takes only a few seconds, less time, in fact, than it takes to read the same set of patterns from disc into the vector processor memory!

The spline algorithm is the heart of a powerful research tool which permits testing of any sort of frequency warping for speaker normalization. To date we have experimented only with one-parameter frequency-scaling transformations,
but the software we have developed could serve also to implement the more complicated types of transformations which other researchers have described[2,3].

2. E.P. Neuberg, "Frequency-Axis Warping to Improve Automatic Word Recognition" (included in September 1980 report)
3. H. Matsumoto and H. Wakita, "Frequency Warping for Nonuniform Talker Normalization" (included in September 1980 report).

C. Implementation and Results for Isolated Words

For the standard isolated-word vocabulary of digits, "yes", and "no", we generated sets of frequency-scaled versions of standard reference patterns, ranging from patterns scaled down by 15 percent, appropriate for recognizing

speakers with unusually long vocal tracts, to patterns scaled up by 30 percent, appropriate for recognizing speakers with unusually short vocal tracts. To say that a reference pattern is "scaled up by 30 percent" means that the parameters of this pattern at frequency f equal the interpolated value of the standard parameters at frequency f/1.30.

From recognition experiments with many different frequency-scaled patterns simultaneously active, we found that misrecognitions due to the large number of patterns frequently prevent any significant improvement in net recognition rate. To take advantage of frequency-scaled patterns, it is desirable to determine, for each speaker, the most appropriately scaled pattern or patterns, and to use only these for recognition. Our most effective technique for calculating the optimum shifted pattern is to recognize a few words with many scaled patterns, in order to compute the average recognition score, then, by carrying out a least-squares fit of a parabola to these average scores, to determine what frequency scaling would produce the lowest average score. This procedure is illustrated in Figure 3 below, which shows a situation where the optimum patterns would be scaled up by about 8% in frequency form the standard patterns. This technique yields a normalization parameter for each speaker which turns out to be quite independent of the particular set of scaled patterns used in its construction. Patterns spaced by 3% or 4% yielded essentially the same parameters as those spaced by 5%.

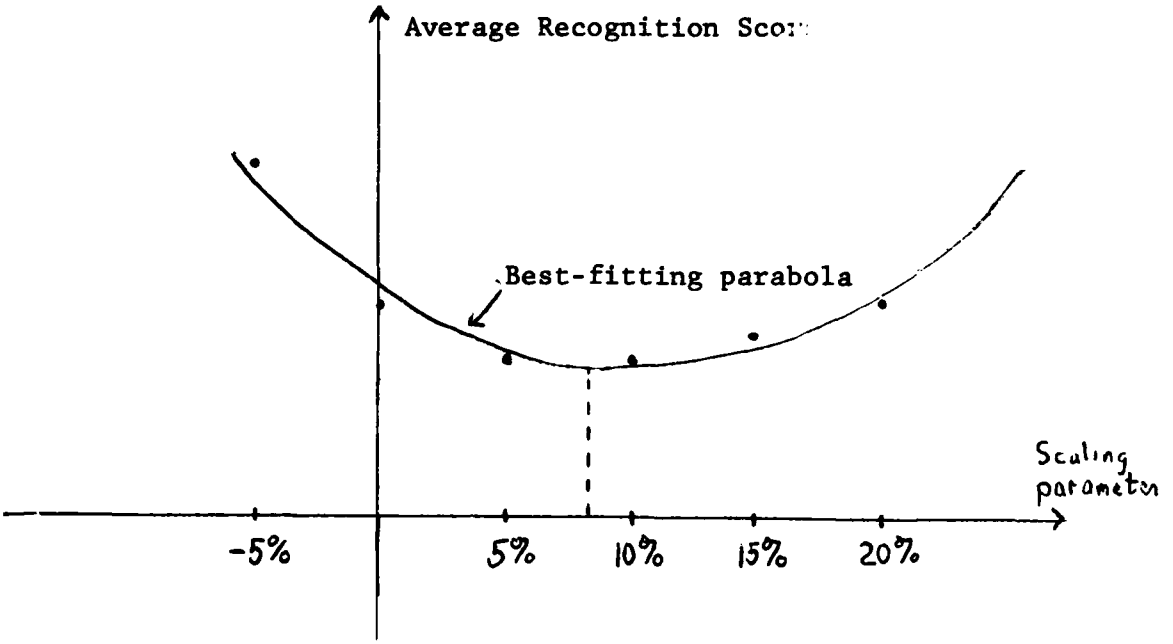


Figure 3 - Estimating a Normalization Parameter

Having calculated a single normalization parameter for each speaker, we next used that parameter to transform the speaker's utterances, thereby creating a set of normalized input data, from which we then generated normalized reference patterns. We also conducted numerous experiments in normalization of speaker-dependent patterns, in which we calculated the normalization for one speaker relative to another, then recognized by using a set of optimally shifted patterns.

Speaker-independent isolated-word experiments were performed using a database of 1910 utterances from 90 speakers, 82 males and 8 females. We carried out the following sequence of tests to assess the effect of frequency-scaling the reference patterns, of frequency-scaling the input acoustic

parameters, and of combining the two approaches.

1. Reference patterns were generated by averaging all utterances. This single set of patterns was then used to recognize all utterances.

   Result: 1767 correct (92.5%)

2. Using a set of frequency-scaled patterns, ranging from down 12% to up 28%, a normalization parameter was computed for each speaker, then the reference patterns closest to the computed best scaling factor were used for recognition.

   Result: 1809 correct (94.7%)

3. Test 2 was repeated, but the recognition scores for the correct word, rather than those for the best-scoring word (right or wrong) were used in computing normalization parameters.

   Result: 1816 correct (95.1%)

4. The normalization parameters computed in test 2 were used to transform the input data from each speaker. Recognition of the normalized input data was carried out with only the original unshifted patterns.

   Result: 1811 correct (95.3%)

5. From the normalized input data of test 4, new reference patterns were generated, and these patterns were used to recognize the original unnormalized utterances.

   Result: 1751 correct (91.7%)

6. Several shifted versions of the normalized patterns

from Test 5 were used for recognition of the unnormalized utterances, as in Test 2

Results: 1813 correct (94.9%)

7. Test 6 was repeated in the manner of Test 3, with knowledge of the true word being used in computing normalization parameters.

   Results: 1821 correct (95.3%)

8. The normalized patterns of Test 5 were used to recognize the normalized data from which they were generated.

   Results: 1817 correct (95.1%)

9. Test 8 was repeated, with knowledge of the true word used in computing normalization parameters.

   Result: 1830 correct (95.8%)

During most of the preceding tests, the number of utterances used to compute normalization parameters was varied. Typically, use of more than twelve utterances produced no significant improvement, and it was possible to use as few as six utterances with a loss of only 0.1% in recognition accuracy.

Detailed analysis of the normalization tests revealed that most of the improvement resulted for speakers whose calculated normalization parameter was greater than five percent (females) plus a few males with parameters of -10%. To assess the improvement in such cases, we conducted several tests in which reference patterns were made from speakers of one sex only, as follows:

10. Patterns made from 82 males were used to recognize 172 utterances of 8 females.

    Result: 106 correct (62.0%)

11. Seven scaled versions of these male patterns were used for recognition of females.

    Result: 147 correct (85.5%)

12. The original unscaled male patterns were used to recognize normalized female utterances.

    Result: 144 correct (83.7%)

13. Patterns made from 8 females were used to recognize 1738 male utterances.

    Result: 1112 correct (64.0%)

14. Six scaled versions of these female patterns were used for recognition of males.

    Result: 1418 correct (81.6%)

15. Patterns made from normalized female utterances were used to recognize male utterances.

    Result: 1429 correct (82.2%)

In a final series of experiments with isolated digits, reference patterns were generated from utterances of individual speakers, then frequency-scaled versions of these patterns were used to recognize other speakers. Striking improvement was obtained in all cases where the calculated optimal frequency shift was 10 percent or more. The most remarkable results are an improvement from 59.2% to 93.6% recognition which resulted from a 25% frequency shift and an improvement from 30.8% to 77.4% when the same shift was

applied to reference patterns made from a single utterance of each digit. The experimental procedure and complete results are in the Appendix. The results are given in Tables 3 and 4.

What we have learned from these experiments with isolated words may be summarized as follows:

1. Normalization by uniform frequency scaling, in a diverse male-female database, can eliminate roughly half of all recognition errors. In cases where patterns made from speakers of one sex are used to recognize speakers of the other sex, or where speaker-dependent patterns are involved, the improvement is somewhat greater.
2. Essentially all improvement comes from using shifts of more than five percent, and most of the possible improvement is obtained if patterns for shifts which are multiples of five percent are available.
3. Recognition of six to twelve utterances is adequate for computing a single normalization parameter. If the identity of these utterances is known, a slight additional improvement results.
4. Normalizing the input data has essentially the same effect as normalizing the reference patterns.
5. Reference patterns made from normalized data are superior to those made from unnormalized data, but only if normalization is used in the recognition process.

D. Implementation and Results for Keyword Recognition

As a preliminary step in applying speaker normalization

Run	M 3	M 4	M 5	F 6	F 2	F 1
M 3	99.1 99.0 99.2 99.1 0					
M 4	84.8 (1) 93.7 (2) 96.0 (3) 96.4 (4) +10	99.5 99.4 99.4 99.5 0				
M 5	82.0 88.8 95.0 97.5 +10	98.4 98.1 98.6 98.4 0	99.9 100.0 100.0 99.9 0			
F 6	67.1 82.1 87.3 87.2 +15	92.0 92.3 94.9 93.8 +10	82.4 85.9 88.5 90.7 +10	99.1 99.3 99.1 99.1 0		
F 2	60.4 78.5 79.6 79.6 +20	59.4 72.6 73.7 73.7 +10	54.3 67.4 72.0 71.6 +10	79.0 77.9 78.9 79.0 0	99.6 99.5 99.5 99.6 0	
F 1	59.2 90.5 93.5 93.6 +25	63.8 86.5 83.8 85.1 +10	62.8 89.9 92.6 93.1 +20	89.1 90.4 92.0 92.7 +5	76.2 72.2 73.7 76.7 +10	99.7 99.7 99.7 99.7 0

(1) Noshift percent recognition
(2) All shifts
(3) Best 2 shifts
(4) Best shift

Table 3

Frequency Shift Experiments Means only SD = 25

Run	M 3	M 4	M 5	F 6	F 2	F 1
M 3	92.6 (1) 90.2 (2) 92.6 (3) 0					
M 4	54.2 67.3 69.7 +10	95.4 93.2 95.4 0				
M 5	64.8 79.3 81.9 +10	93.0 93.5 93.0 0	99.3 98.2 99.3 0			
F 6	43.8 69.1 75.7 +20	76.8 81.7 82.5 +5	67.4 78.3 74.7 +10	95.1 93.2 95.1 0		
F 2	41.2 78.9 71.5 +20	54.5 66.6 70.6 +10	38.9 54.3 58.9 +15	57.5 56.7 61.1 +5	96.3 94.6 96.3 +0	
F 1	30.8 62.2 77.4 +25	48.7 67.9 75.5 +15	51.1 73.0 65.2 +15	79.8 84.4 84.7 +5	70.7 68.9 74.6 +10	82.9 80.2 82.0 0

(1) Noshift percent recognition

(2) All shifts

(3) Best shift

Table 4

Frequency Shift Experiment ---Single Utterance Templates

to keyword recognition, we transferred picked-pattern files for linearly-segmented keywords to the VAX and applied the isolated-word normalization techniques to them. The results were similar to those for digits. When reference data and test data were all from speakers of one sex, normalization produced no significant improvement, but when English female speakers were recognized by using frequency-scaled English male patterns, recognition increased to 88.3% from the 78.5% obtained with unshifted male patterns. The females in the database almost all had a normalization parameter close to fifteen percent.

To implement speaker normalization in the keyword algorithm, three new commands have been added to the system. One of these simply extracts each reference pattern in turn from one of the reference pattern files in the vector computer, frequency-scales it up or down by a specified percentage, and copies the result into the other reference pattern file, which is then used for recognition. The second command permits the operator to create, in the secondary word array file, word arrays for frequency-scaled versions of existing words. The final command automatically generates frequency-scaled patterns in accordance with the specifications in the secondary word array file. Because of the limit of twenty active sets of patterns, it is impossible to have multiple patterns for all vocabulary words at once, but it is straightforward, for example, to generate word arrays and patterns with shifts of -5,0,5,10, and 15 percent for each of four existing vocabulary words.

To test the effectiveness of speaker normalization in keyword recognition, we have run the standard recognition tests using the design database of English female speakers as test data to be recognized by various frequency-scaled versions of the English male reference patterns. The reference patterns and word arrays used were those developed under the previous contract; we have not yet repeated the test with the recently created improved reference patterns and word arrays.

Table V summarizes the detection rates, for 5, 10,15 and 20 false alarms per word, when various normalization parameters were used. At the end of the table, for comparison, are the results obtained by using patterns generated directly from the female design data and the earlier results for recognition of male design data and test data (Test 1 of Chapter II).

	FALSE ALARMS PER WORD			
	5	10	15	20
Male patterns, unshifted	15%	22%	26%	29%
Male patterns, up 6%	27%	35%	39%	43%
Male patterns, up 9%	32%	39%	45%	48%
Male patterns, up 12%	34%	43%	48%	51%
Male patterns, up 15%	37%	45%	50%	53%
Male patterns, up 18%	36%	47%	52%	53%
Male patterns, up 21%	36%	46%	50%	53%
Female patterns, unshifted	62%	68%	73%	75%
Recognition of male design data	60%	68%	72%	74%
Recognition of male test data	39%	48%	51%	55%

Table 5: Test Results for Female Speaker with Frequency-Scaled Male Reference Patterns.

It is clear from these results that frequency-scaling produces striking improvement in recognition results when female speech is recognized using male reference patterns. The best normalizations, 15 and 18 percent, are the same as calculated from the isolated-word experiments on these keyword data, and they both lead roughly to a doubling of the detection rate. What is not clear is whether normalized male patterns are in any sense an adequate substitute for female patterns. The only female patterns available were generated from the same data that was used for testing, and they should be expected to perform better than any patterns made from independent data. The shifted male patterns actually performed almost as well in

recognizing independent female test data as did the unshifted male patterns in recognizing independent "Stonehenge" test data, but the "Stonehenge" database is intrinsically harder to recognize than the design database

The full algorithm for computing normalization parameters has not yet been transferred to the keyword system, but considerable progress has been made. The capability of the keyword system to generate several frequency-scaled versions of a word's reference pattern and to perform simultaneous detection of a word with several different patterns has been verified, and inspection of a small amount of test data indicates that the recognition scores will be adequate for computing normalization parameters. Furthermore, the fact that the optimal frequency shift for females as a group turned out to be in the same 15-18 percent range for keywords as well as for isolated words suggests that the isolated-word techniques ought to carry over well to keyword spotting. What is missing is the software that will automatically convert detection scores for shifted patterns into normalization parameters. Once this software has been designed, it will become possible for the keyword system automatically to generate appropriate shifted reference patterns for a new speaker after it has detected a few keywords.

E. Conclusions and Suggestions for Future Research

Frequency scaling has proved effective for improving recognition of female speakers from male patterns, although it appears unlikely that any normalization technique will generate patterns that are as effective as ones made from a large female

database. The straightforward scaling scheme we have used, while less elegant and probably less powerful than some described in the literature, has one great advantage: it involves only one parameter, which can be reliably estimated from a small number of unverified detections.

Much of the effort expended on normalization has gone into writing the code to implement the spline algorithm, probably the most complex computation ever undertaken on the Verbex vector processor. Now that the capability for rapid transformation of reference patterns exists, the following lines of research can be pursued:

1. Automatic computation of normalization parameters. This could lead to a system which initially uses many frequency-scaled patterns for a small set of words then, as the normalization for the current speaker became better known, uses fewer and fewer alternate patterns for more and more words until finally it is operating with one optimally transformed pattern for each word.

2. Design of better normalization transformations

Studies of frequency warping [2,3] describe transformations which perform significantly better than uniform scaling in some contexts. Once such a transformation has been developed, and its parameters have been incorporated into the spline algorithm, it will operate as fast as the simple transformations now in use. Storing carefully designed transformations for speakers who must frequently be recognized may prove to be an attractive alternative to generating and storing may sets of speaker-dependent patterns.

3. Normalization of initial patterns for training. The on-line

training procedure described in Chapter III uses an initial template derived from isolated words spoken by a cooperative speaker. If it is desired to create patterns for a speaker who cannot be asked to speak words in isolation, then a frequency-scaled version of the patterns generated from the operator's isolated utterances is likely to be the best starting point. The capability to do this now exists on the keyword system but has not yet been tested. The obvious challenge is for a male operator to generate adequate reference patterns for a female for whom only unlabeled, continuous speech is available.

CHAPTER V

Keyword Performance

Test results are presented in this chapter for the government-furnished "Stonehenge" data in a format which permits direct comparison with Section III of the previous RADC technical report.[2]

Q2 Telephone Quality, Stonehenge Male Test Data

		<1 FA in 1.19 hrs. <0.8 FA/hr		3FA in 1.19 hr. 2.5 FA/hr		10 FA in 1.19 hr. 8.4 FA/hr	
1.	Boonsboro	25/101	25	42/101	42	60/101	59
2.	Chester	7/86	8	25/86	29	56/86	65
3.	Conway	27/43	63	38/43	88	40/43	93
4.	Interstate	3/58	5	12/58	21	21/58	36
5.	Look	0/34	0	0/34	0	1/34	3
6.	Middleton	20/62	32	34/62	55	40/62	65
7.	Minus	2/45	4	13/45	29	20/45	44
8.	Mountain	7/63	11	10/63	16	13/63	21
9.	Primary	6/41	15	31/41	76	36/41	88
10.	Retrace	11/36	31	21/36	58	32/36	89
11.	Road	6/63	10	13/63	21	27/63	43
12.	Secondary	7/21	33	10/21	48	13/21	62
3.	Sheffield	14/31	45	17/31	55	22/31	71
14.	Springfield	15/29	52	17/29	59	21/29	72
15.	Thicket	3/27	11	8/27	30	8/27	30
16.	Track	6/54	11	10/54	19	31/54	57
17.	Want	0/48	0	1/48	2	1/48	2
18.	Waterloo	4/52	8	9/52	17	16/52	31
19.	Westchester	23/49	47	35/49	71	39/49	80
20.	Backtrack	8/12	67	10/12	83	11/12	92
			24%		41%		55%

Table 6: Results for Q2 English Males, 20 Words

Tapes 1, 2, 3, 4, 5, 6, 7, 8

Keyword - Spotting - Verbex

Stonehenge Tapes 1, 2, 3, 4, 5, 6, 7, 8

	Q2 Males		English 1.19 hrs.
	OFA 1FA	3FA/1.19 hrs. 2.25 FA/Hr.	10FA/1.19 hrs. 8.4 FA/Hr.
Boonsboro	25	42	59
Conway	63	88	93
Interstate	5	21	36
Look	0	0	3
Middleton	32	55	65
Mountain	11	16	21
Primary	15	76	88
Secondary	33	48	62
Sheffield	45	55	71
Westchester	47	71	80
Males only avg.	28%	47%	58%
Males only without Look, Mountain	33%	57%	69%

Table 7: Results for Q2 English Males, 10 Words

Keyword - Spotting - Verbex

Stonhenge Tapes 1, 2, 3, 4, 5, 7, 8, 9

		Q3 Males				English 1.05 hrs.	
		OFA 1FA/1.05hr.		3FA/1.05 hrs. 2.9 FA/hr.		10/1.05 hrs. 9.5 FA/hr.	
1	Boonsboro	14/91	15	27/91	29	43/91	47
2	Chester	18/76	23	31/76	40	39/76	51
3	Conway	19/37	51	24/37	64	28/37	75
4	Interstate	2/46	4	7/46	15	13/46	28
5	Look	0/38	0	0/38	0	2/38	5
6	Middleton	17/58	29	26/58	44	30/58	51
7	Minus	7/45	15	11/45	24	19/45	42
8	Mountain	2/57	3	5/57	8	9/57	15
9	Primary	2/36	5	21/36	58	26/36	72
10	Retrace	3/37	8	25/37	67	29/37	78
11	Road	0/54	0	5/54	9	7/54	12
12	Secondary	2/19	10	6/19	31	11/19	57
13	Sheffield	4/25	16	17/25	68	17/25	68
14	Springfield	0/21	0	6/21	28	9/21	42
15	Thicket	1/21	4	3/21	14	6/21	28
16	Track	8/49	16	9/49	18	24/49	48
17	Want	0/37	0	1/37	2	4/37	10
18	Waterloo	1/43	2	8/43	18	12/43	27
19	Westchester	11/46	23	27/46	58	31/46	67
20	Backtrack	4/12	33	6/12	50	9/12	75
			13%		32%		45%

Table 8: Results for Q3 English Males, 20 Words

Keyword - Spotting - Verbex

Stonehenge Tapes 1, 2, 3, 4, 5, 7, 8, 9

Q3 Males

	0FA 1FA/1.05 hr	3FA/1.05 HRS. 2.9 FA/hr	10FA/1.05 HRS. 9.5 FA/hr
Boonsboro	15	29	47
Conway	51	64	75
Interstate	4	15	28
Look	0	0	5
Middleton	29	44	51
Mountain	3	8	15
Primary	5	58	72
Secondary	10	31	57
Sheffield	16	68	68
Westchester	23	58	67
Males only avg.	16%	38%	48%
Males only without Look, Mountain	19%	46%	58%

Table 9: Results for Q3 English Males, 10 Words

Keyword - Spotting - Verbex

Stonehenge Tapes 1, 3, 7, 8 — English
Q2 Females — 0.35 hour

	OFA <1FA/0.35hr.		2FA/0.35 hrs. 5.7FA/hr		5FA/0.35 hrs. 14.3FA/hr.		
1	5/30	17	10/30	33	14/30	47	Boonsboro
2	5/26	19	6/26	23	7/26	27	Chester
3	3/19	16	7/19	37	7/19	37	Conway
4	0/21	0	0/21	0	1/21	5	Interstate
5	0/2	0	0/2	0	1/2	50	Look
6	0/25	0	3/25	12	6/25	24	Middleton
7	3/24	13	9/24	38	15/24	63	Minus
8	0/21	0	5/21	24	10/21	48	Mountain
9	5/23	22	13/23	57	15/23	65	Primary
10	1/2	50	2/2	100	2/2	100	Retrace
11	2/29	7	3/29	10	3/29	10	Road
12	5/16	31	8/16	50	8/16	50	Secondary
13	1/7	14	1/7	14	1/7	14	Sheffield
14	1/5	20	1/5	20	1/5	20	Springfield
15	0/12	0	0/12	0	0/12	0	Thicket
16	4/22	18	4/22	18	9/22	41	Track
17	0/8	0	1/8	13	1/8	13	Wave
18	6/12	50	8/12	67	8/12	67	Waterloo
19	4/13	31	4/13	31	11/13	85	Westchester
20	1/1	100	1/1	100	1/1	100	Backtrack
		20%		32%		43%	

Table 10: Q2 English Females (with 1st pass female patterns) 20 Words

Keyword - Spotting - Verbex

Stonehenge Tapes 1, 3, 7, 8, 9 English
Q3 Females 0.404 hours

	OFA < 1FA/0.4 hr.		2FA/0.4 hrs. 5.0FA/hr.		5FA/0.4 hrs. 12.5FA/hr.		
1	2/39	5	5/39	13	6/39	15	Boonsboro
2	3/28	11	3/28	11	7/38	25	Chester
3	8/22	36	10/22	45	11/22	50	Conway
4	0/28	0	3/28	11	3/28	11	Interstate
5	0/5	0	0/5	0	0/5	0	Look
6	10/32	31	10/32	31	12/32	38	Middleton
7	7/27	26	9/27	33	10/27	37	Minus
8	5/26	19	8/26	31	8/26	31	Mountain
9	3/25	12	7/25	28	8/25	32	Primary
10	0/6	0	2/6	33	2/6	33	Retrace
11	2/31	6	2/31	6	2/31	6	Road
12	5/18	28	6/18	33	7/18	39	Secondary
13	0/10	0	1/10	10	1/10	10	Sheffield
14	1/10	10	1/10	10	1/10	10	Springfield
15	0/13	0	0/13	0	1/13	8	Thicket
16	3/26	12	3/26	12	4/26	15	Track
17	0/9	0	0/9	0	0/9	0	Wave
18	2/19	11	8/19	42	9/19	47	Waterloo
19	2/21	10	8/21	38	13/21	62	Westchester
20	1/1	100	1/1	100	1/1	100	Backtrack
		16%		24%		28%	

Table 11: Results for Q3 English Females (with 2nd pass female patterns) 20 Words

Appendix

Speaker Normalization for Single-Speaker Recognition of Digits

Experimental results obtained for the recognition of single speaker random digits against shifted patterns of other speakers in the RAND 1 - RAND 6 database are given in Table 3. All pairs were run for which the best shift was found to be positive or zero, thereby obtaining one comparison for each pair of speakers. Speakers M3, M4, and M5 are males; F1, F2, and F6 are females. Four recognition scores are given for each run as indicated at the bottom of the table. The best shift (percent) is given in the lower right of each block. Note the substantial increase in performance by use of shifted patterns, especially when the shift percentage is large, roughly corresponding to a large difference in vocal tract lengths. It appears that much of the remaining error is due to time misalignment, arising due to use of the isolated word algorithm, which should be much less of a problem in the keyword algorithm.

A similar experiment was done by using a single utterance of each word as a reference pattern. A possible approach to the problem of automatic generation of keyword patterns is to start from a single utterance and collect additional examples by a semiautomatic procedure using the single-word pattern in a dynamic programming algorithm. The algorithm would aid in both word labeling and phone segmentation, and would help to realize better time alignment in the generation of patterns. Frequency shifting would be vital in the above task to provide high accuracy on speakers whose vocal tract lengths differ substantially from the initial prototype

speaker.

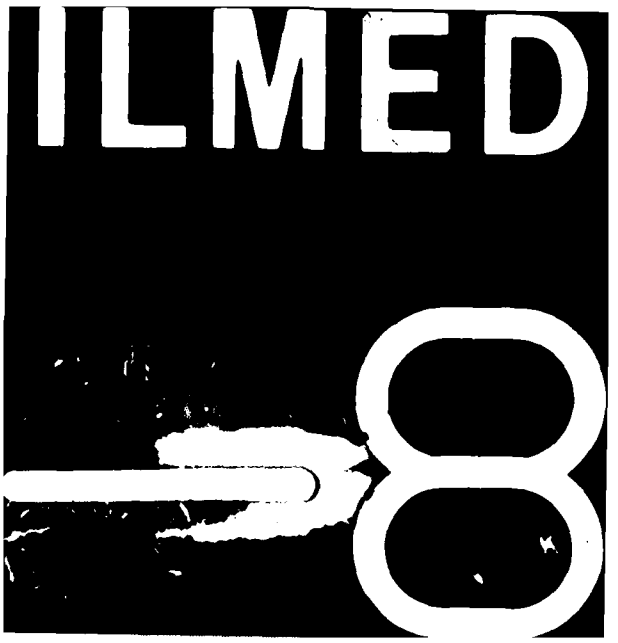
Table 3 gives results obtained on single utterance patterns in a form similar to Table 4. The value of frequency shifting is readily apparent in these results, although total performance is not high. Much of the remaining error can be attributed to time misalignment due to the isolated word algorithm, which would not be present in the keyword application.

BIBLIOGRAPHY

1. Moshier, S.L. , Bahler, L., "Keyword Operational Analysis" RADC TR 81-228, Aug 1981

2. Neuberg, E.P., "Frequency-Axis Warping to Improve Automatic Word Recognition" Proceedings of IEEE-ASSP, 1980.

3. Matsumato, H. and Wakita, H., "Frequency Warping for Nonuniform Talker Normalization" - Proceedings of IEEE-ASSP, 1979


ILMED
8